The influence of experimental design on the magnitude of the effect size -peer tutoring for elementary, middle and high school settings: A meta-analysis

Zeneli, M., Thurston, A., & Roseth, C. (2016). The influence of experimental design on the magnitude of the effect size -peer tutoring for elementary, middle and high school settings: A meta-analysis. *International Journal of Educational Research*, *76*, 211. https://doi.org/10.1016/j.ijer.2015.11.010

Published in:
International Journal of Educational Research

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

The influence of experimental design on the magnitude of the effect size - peer tutoring for elementary, middle and high school settings: A meta-analysis

Mirjan Zeneli
Correspondence: School Of Education, Durham University, Stockton Rd, Durham, County Durham, DH1 1TA, UK. *Email*: mirjan.zeneli@durham.ac.uk , *Tel*: 078 078 5 5353; *fax*: +00 44 (0) 191 334 8311.

Allen Thurston
a.thurston@qub.ac.uk
Centre for effective Education,
College of Education,
Queens University, Belfast, UK, BT7 1HL

Cary Roseth
croseth@msu.edu
Michigan State University
Michigan
USA
MI48824

The influence of experimental design on the magnitude of the effect size - peer tutoring for elementary, middle and high school settings: A meta-analysis

Mirjan Zeneli*, Allen Thurston & Cary Roseth

Abstract

A meta-analysis was undertaken on a form of cooperative learning, peer tutoring. The effects of experimental design on outcomes were explored, as measured by *Effect Size (ES)*. Forty one articles with 49 studies were included in the meta-analysis. ES were positive for peer tutoring with highest effects in elementary, reciprocal role tutoring conducted with low socio-economic class students with high ethnic minorities. ES was influenced by experimental design including design, matching of population samples at pre-test and selection of attainment measures. The implications for future meta-analyses and educational research design in peer tutoring and more broadly are explored.

***Correspondence**: School of Education, Durham University, Durham, UK.

Email: mirjan.zeneli@durham.ac.uk, *Tel*: 078 078 5 5353

1. Introduction

Previous meta-analyses have concluded that peer tutoring had a positive impact on academic achievement (e.g., Cohen, Kulik & Kulik, 1982; Cook, Scruggs, Mastropieri, & Casto, 1985; Leung, in press 2014; Mathes & Fuchs, 1994; Rohrbeck, Ginsburg-Block, Fantuzzo & Miller, 2003). However, it remains unclear whether the type of research design used to study treatment effects may bias the magnitude of effect sizes (ES). This is a problem, since any widespread methodological features that bias treatment effects in primary studies will also bias the mean effect size estimates reported in meta-analyses (Lipsey & Wilson, 1993).

For peer tutoring, a mix of mean treatment effect sizes have been reported in previous meta-analyses: Table 1 indicates that the overall mean ES (fixed effect) found in each meta-analysis ranged from 0.26 to 0.75, and the number of studies on which these figures were based ranged from 11 to 72.

Of course, what remains unclear about the range of mean ES for peer tutoring, and other interventions, is whether some methodological artefact of the primary studies accounts for some of the variability. For example, a randomized controlled trial (RCT) might be associated with smaller effect size estimates than single group or quasi-experimental studies because randomization on pre-test data controls factors (e.g., history threats, maturation threats, testing threats, instrumentation threats, mortality, regression threats) that might inflate ES estimates (Campbell & Stanley 1963; Holland, 1986; Trochim, 2012; Tymms, Merrell & Coe, 2008).

Supporting this view recent RCT studies of peer tutoring have reported lower ESs than a study using the same tutoring techniques in studies when randomisation was not used to assign to condition. For example, the Fife Peer Learning project where tutoring in reading and maths was implemented over two years in a RCT reported a mean ES of 0.20-0.25 (Thurston & Topping, 2008; Topping & Thurston, 2008; Tymms, Merrell, Thurston, Andor, Topping & Miller, 2011). The same peer tutoring technique used in matched, rather than randomised, experiment designs has yielded ES of between 0.46 and 1.0 (Thurston, Burns, Topping & Thurston, 2012; Thurston, MacNia & Keenan, 2015). The mean ES of the Fife Peer Learning project was also lower than might have been anticipated based on the meta-analyses of Cohen, Kulic and Kulic (1982).

An important issue with RCTs in Education and in Social Sciences is that currently there are no clear guidelines to aid researchers on the steps they need to take in order to provide coherent and trustworthy research as is the case in medical research with the Consolidated Standards of Reporting Trials (CONSORT). Therefore, we also recommend that future researchers concentrate on developing and implementing a set of rules and guidance for conducting experiments and trials in Social Sciences in general and suggest what features such a guide would require.

2. Literature review: The significance and need for this study

2.1 Review on the impact of research design

Few meta-analysts have examined whether the type of experimental design moderated ES estimates. One of the first papers to look at this issue was Lipsy and Wilson (1993). It was concluded that single case studies had a higher reported ES. Slavin, Lake and Groff's (2009) systematic review on what method works in mathematics, compared reported ESs between a) retrospective matching (ex-post facto design) studies with perspective matching, b) between perspective matching studies with randomised studies, c) and between small and large sample size studies. Their conclusion was that ex post-facto studies showed higher ESs than perspective matching and studies with smaller sample sizes showed larger ESs than studies of larger sample sizes. Nevertheless, they found no difference in reported ESs between RCTs and matched studies. However, both systematic reviews had the primary focus of reporting ES based on heterogeneous interventions.

Lipsy and Wilson (1993) examined differences on reported ESs attributable to research designs by using study areas with quite diverse characteristics.

To date, only Carlson and Schmidt (1999) have compared the reported ESs between two homogeneous groups, single pre-post treatment-step design versus experimental control designs within a single area of training. Their conclusion was that experimental control designs studies yielded lower reported ESs. Nevertheless, Carlson and Schmidt concentrated mostly on the effect produced by different formulas in calculating ESs rather than experimental designs themselves.

There are several features of different experimental designs that might influence ES estimates. There is a strong acknowledgment among academics that randomisation is unsatisfactory in assuring equalisation between treatment and control samples (Slavin, 2008b: Lachin, Matts & Wei, 1988). It has been recommended that at least two further checks need to take place, namely: a) that strict or blocked-randomisation is in place (Lachin, et al, 1988), b) that the sample size in each group is large, so that when randomisation takes place probability theory ensures that the groups are allocated similar participants (Slavin, 2008b) or that the interventions, and of course therefore units of assessment, are clustered, i.e. take place at the school level (Thurston, 2008).

Another important topic in the planned meta-analysis will be whether the design and origin of assessment instruments make a difference to the reported ESs.

Ruiz-Primo, Shavelson, Hamilton, and Klein, (2002) conducted a RCT using two different tests: they reported a correlation between distal measurement (standardised state instrument) and low ES, arguing that the main theoretical justification for such phenomenon is the notion that standardised state instruments are broader in content than researcher-made instruments, which tend to focus on specific topics taught in the experimental groups (Herman, et al, 2006). Slavin and Maden (2008, 2011) concluded that research instruments inherent to experimental treatments showed higher ESs. The study, however, also examined research which concentrated on heterogeneous interventions.

This review aims to contribute to the three-way debate, between: a) Those who argue that meta-analysis should use strict inclusion criteria (Oakley, 2006, Slavin 2008), b) those who suggest weaker designs to be treated with equality in evidence-based research (Hammersley, 1997; Morrison, 2001), and c) those who recommend the inclusion of weaker designs, however, make use of some form of statistical model that accounts for the differing characteristics (Rubin, 1993). By concentrating on instrumentation the paper also aims to shed light on the role of instruments and measures.

2.2 Review on the impact of peer tutoring

Previous meta-analysis of peer tutoring that have concentrated on elementary and high school students of ages 4-18 have included those of Cohen, et al, (1982), Bowman, et al, (2013). Cohen,

et al's, (1982) review was based on three inclusion criteria:1) peer tutoring to be conducted at school, 2) studies needed to contain the necessary coefficients in order to calculate the ES, and 3) the studies needed do to have a strong methodology. The limited number of inclusion criteria did not ensure sample homogeneity.

The closest meta-analysis paper which comes to replicate the Cohen, et al (1982), in terms of sample population age 4-18 is that by Bowmen-Perot, Davis, Vannest, Williams, Greenwood, Parker (2013). This paper, however, concentrated on single-case designs.

The latest meta-analysis in peer tutoring is that by Leung (2014). The paper includes studies; i) less than 6 weeks, ii) peer tutoring interventions on special education, and ii) studies with participants with ages over 18 years of age. Whereas the last element would influence the overall effect size, the first and the second inclusion criteria would also have an influence on the average peer tutoring ES on elementary and high school students. Meta-analysis report that peer tutoring interventions on disabled students provide higher effect sizes 0.76 compared to 0.65, which may have skewed results and conclusions (Bowman-Perrott, et al, 2013).

Table 1 below provides a review which supports the need for another meta-analysis in peer tutoring for students of school age 4-18:

3. Aim and Objectives:

The main aim of this study is to explore the impact of experimental design on ES outcomes reported by peer tutoring studies. The study was focussed on studies reported to take place in , elementary and high school students. Four main questions were addressed in the study:

i). What role does experimental design play in reported effect size by studies on peer tutoring?

ii). How do effect sizes reported from author developed assessment instruments differ when compared to instruments developed independently of the research team when measuring academic changes in studies of peer tutoring?

iii). What is the overall reported effect size on student attainment for peer tutoring interventions for students aged 4-18 years-old?

iv). How does reported effect size vary with implementation changes during p[eer tutoring in respect of peer tutoring structure, training, and population characteristics?

Table 1

Meta-analysis in peer tutoring

Meta-analysis	*Population Characteristics*	*Academic Fixed ES*
Cohen, et al (1982)	Peer Tutoring with 4-18 year olds, on mathematics and reading. Totalling 52 studies.	0.40
Cook, Scruggs, Mastropieri, and Casto (1986)	Peer tutoring on disabled students. Mathematics and Reading. Totalling19 studies.	0.59
Mathes & Fuchs, (1994)	Peer tutoring on disabled students, reading. Cross age, same age, fixed and reciprocal. 11 studies.	0.42
Rohrbeck, et al, (2003)	Pairs and small groups, ages 5.65 to 11.50. Including studies from less than 6 weeks. 40 out of 90 studies were peer tutoring.	0.33
Ginzburg-Block, Rohrbeck, and Fantuzzo (2006)	Peer Assisted Learning studies, as oppose to peer tutoring in general, on grades 1-6 students. Totalling, 26 academic outcome studies.	0.35
Bowman-Perrott, et al (2013)	Academic benefits of peer tutoring on grades 1-12. Totalling 26 single case research studies.	0.75

Leung (2014)	Peer tutoring elementary, high and university students including special needs populations. 72 studies, including studies with less than six weeks.	0.26

4. Methods

4.1 Peer Tutoring Definition

Peer tutoring is a specific form of cooperative learning. It takes place in pairs or triads and tends to have clear patterns for interaction. Peer tutoring has been described as a "*form of peer learning. It generally involves one student teaching another where pairs are typically of differing academic standing and sometimes differing ages*" (Tymms, et al, 2011, p267). Peer Tutoring (PT) takes a number of forms: 'Peer-Assisted Learning Strategies' (PALS), 'Cross-Age Tutoring' (CAT) and 'Reciprocal Peer-Tutoring' (RPT) (Topping & Ehly, 1998).

4.2 Sampling Procedures and Criteria

Every effort was made to locate all relevant studies and to minimize any systematic data exclusion. Psych INFO and ERIC online databases were searched using keywords: *peer tutoring, peers, peer relations, cross-age tutoring, cooperative learning, transfer learning, cross-age learning, peer learning, reciprocal peer learning, reciprocal peer tutoring, and peer assisted learning*.[1]

To be included in this meta-analysis, studies had to meet several criteria:

1) Studies included in the meta-analysis employed a form of peer tutoring, defined as one/two students teaching one/two others working in a pair or triad.

2) Studies had to be published in English in a peer-reviewed journal from January 1965 until December 2014.

[1] A complete list of articles identified using these procedures can be obtained upon request.

3) Studies comparing one form of peer tutoring to another were excluded because relative effects would not be comparable to effect sizes based on control-group or pre-test comparisons.

4) Studies had to (a) measure academic achievement, (b) involve students ranging in age from 4 to 18 years of age (elementary to high school age), and (c) involve an intervention lasting at least or longer than 6 weeks. Studies exclusively involving students with learning disabilities or studies in which virtual (e.g., computer) peers provided tutoring were excluded.

4.4 Dependent and Independent Variables

The independent variable was research design. A study was categorized as a *pre-post* design if there was no control group. A study was categorized as a *quasi-experimental design* if there was a control group, but no randomization or matching to conditions. A study was categorized as *RCT/1 matched* if the experimental and control groups were matched on only one of the following criteria: (a) previous test performance, (b) socio-economic status/free school meal, (c) ethnic background of the participants, (d) teacher qualifications/length of teaching, or (e) school type. And a study was categorized as a *RCT/2+ matched* if the control and the experimental groups were matched on two or more of the previously listed criteria a-e.

The dependent variable was academic achievement, defined as performance on a learning task. Measures of academic achievement included the Test of Written Language (TOWL-3), Writing Quality and Length (WQL), Academic Information Management System (AIMSweb), and the quality and accuracy of answers on various questions.

The procedures for study characteristics, coder reliability, effect size estimation and data integration such as, calculating average effect sizes, shifting unit of analysis and testing moderator effects, are reported in appendix A and B.

5. Results

5.1 Achieved Sample

The sampling procedures described above initially yielded 13,023 articles from ERIC, 16,908 articles from PsycINFO, and 404 articles from other sources. Of these, 178 unique articles

reported quantitative data on peer tutoring. Applying the inclusion criteria described above reduced the sample to 41 articles reporting 182 separate effect sizes based on 49 separate studies.

5.2 Study Characteristics

Some of the characteristics of the 49 separate studies are listed in Tables 2. Most studies were conducted during the 1990s (33%, *n* = 16), but otherwise fairly evenly distributed across the 1980s (22%, *n* = 11), 2000s (24%, *n* = 11), and 2010s (16%, *n* = 8). Sample size varied by research design, with RCT/2+ match having the largest median, IQR, and range.

Methodology. Similar numbers of studies used a pre-post (35%, *n* = 17), RCT/1 match (35%, *n* = 17), and RCT/2+ match (24%, *n* = 12) designs. The majority conducted a fidelity check (84%, *n* = 36), though the exact distribution varied by research design (e.g., RCT/1 match: *n* = 10 (59%) reported fidelity check, *n* = 7 (41%) did not report). More studies used standardized measures (73%, *n* = 36) than researcher-modified measures (27%, *n* = 13).

Peer tutoring characteristics. Most studies focused on literacy (69%, *n* = 34), followed by mathematics (20%, *n* = 10). There were almost identical numbers of studies that utilized same-age reciprocal peer tutoring (43%, *n* = 21) and cross-age peer tutoring (41%, *n* = 20). Most studies used structured peer tutoring (98%, *n* = 48) and did not use a reward incentive (71%, *n* = 35).

Sample characteristics. The majority of studies were conducted in the United States of America (43%, *n* = 21) and the United Kingdom (35%, *n* = 17). In terms of socio-economic status (SES), measured by free school meals (FSM),most of the studies were mixed (35% FSM, *n*=17), followed by unspecified (29% FSM, *n*=14), and low (27% FSM, *n*=13). On ethnicity, most of the studies (59%, *n*=29) consisted of $\leq$ 50% ethnic group. Most of the studies were conducted on elementary students (88%, *n*=43) and in terms of ability the majority of studies were conducted on mixed (47%, *n*=23) and low abilities (41%, *n*=20).

Group composition. The composition of peer tutoring groups in most studies was heterogeneous with regard to gender (82%, *n* = 40) and ability (82%, *n* = 40).

Intervention characteristics. In the majority of studies, the length of the intervention was 7-12 weeks (41%, *n* = 20), but a significant proportion also lasted 6 weeks (22%, *n* = 11) and 12-

52 weeks (20%, *n* = 10). The mean amount of training – or *training dosage* – was more than 3 sessions (51%, *n* = 22), but this was qualified by the fact that this was unspecified for a substantial number of studies (30%, *n* = 13). Roughly equal numbers of studies were conducted in a general academic settings (49%, *n* = 21) or in intact classrooms (44%, *n* = 19).

Table 2

Study Decade and Methodological and Peer Tutoring Characteristics

	Pre-post	Quasi-exper	RCT/1 match	RCT/2+ match	*n*	%
Total	17	3	17	12	49	100
Sample size						
Median	21.5	16.0	24	119		
IQR	27.5	-	35.7	791		
Range	11, 202	10, 80	8, 268	30, 1246		
Decade published						
1970s	0	2	0	0	2	4
1980s	5	0	5	1	11	22
1990s	3	0	8	5	16	33
2000s	7	1	3	1	12	24
2010s	2	0	1	5	8	16
Fidelity check						
Reported	14	1	10	11	36	73
Not reported	3	2	7	1	13	27
Outcome measure						
Standardised	9	3	10	5	27	55
Researcher modified	6	0	7	6	19	39
Mixed	2	0	0	1	3	6
Subject area						
Mathematics	0	0	2	8	10	20
Literacy	17	3	10	4	34	69
Science	0	0	2	0	2	4
Other	0	0	2	0	2	4
Mixed-core	0	0	1	0	1	2
Type of peer tutoring						
Same-age non-reciprocal	3	0	2	2	7	14

Same-age reciprocal	6	1	7	7	21	43
Cross-age non-reciprocal	7	2	8	3	20	41
Mixed	1	0	0	0		
Structured						
Yes	17	3	16	12	48	98
No	0	0	1	0	1	2
Reward						
Yes	4	0	4	6	14	29
No	13	3	13	6	35	71

5.3 Effect Size Analyses

In all, the 41 articles included in the meta-analysis yielded 182 effect sizes based on 49 separate samples. Averaging the effect sizes across subgroups within each study yielded 49 effect sizes. Tables 3-7 present the findings.

Overall effect of peer tutoring. Firstly, the overall independent effect of peer tutoring on academic achievement for studies using a pre-post design and studies using a control-group design was examined.

Pre-post. For the 13 articles using a pre-post research design, all of the 17 independent effect sizes were in a positive direction. Grubb's test detect one outlier on the right side of the distribution (d = 2.91; Wright & Cleary, 2006[2]), which was Winsorized to the nearest neighbour (d = 1.71) and retained for further analyses. The effects ranged from d = 0.22 to 1.71. The weighted average d was 1.14, $p < .001$, 95% CI [1.04, 1.25], under a fixed-effects (FE) model, and d was 0.92, $p < .001$, 95% CI [0.63, 1.22], under a random-effects (RE) model. The homogeneity statistic for the FE model was also significant, $Q(16) = 104.35$, $p < .001$, $I^2 = 84.67$, indicating that the amount of variance in effect sizes was greater than would be expected from sampling error alone. This finding supports the use of the RE model and the examination of moderators that might influence the effect size distribution.

[2] In the study the ES was lower, as the authors used a different method for calculating the effect size.

To test for possible bias (i.e., missing values within the distribution of effect sizes), trim-and-fill analyses were conducted using both FE and RE models (Borenstein et al., 2005). We found evidence that 8 effect sizes might have been missing from the *right* side of the distribution using a FE model, and imputing these values increased the estimated mean effect to d = 1.49 (95% CI = 1.39, 1.58) under FE and d = 1.42 (95% CI = 1.09, 1.75) under RE. Under the RE model, there was no evidence that any effect sizes might have been missing from either side of the distribution. These findings suggest that the estimated average weighted effect size was somewhat sensitive to missing effect sizes under the FE model but not under the RE model. However, even when controlling for possible bias, the estimated effect size ranged from 0.93 to 1.49. For studies using a pre-post design, this strengthens confidence that the effect of peer tutoring on academic achievement is positive, significantly different from zero, and likely to be large.

Control-group studies. For the 28 articles using a control-group design, only two of the 32 independent effect sizes were in a negative direction and the Grubb's test did not detect any outliers. The effects ranged from d = -0.05 to 1.45. The weighted average d was 0.42, p < .001, 95% CI [0.34, 0.51], under a FE model, and 0.51, p < .001, 95% CI [0.35, 0.67], under a RE model. The homogeneity statistic for the FE model was also significant, $Q(31) = 78.43$, p < .001, I^2 = 60.47, indicating that the amount of variance in effect sizes was greater than would be expected from sampling error alone. This finding supports the use of the RE model and the examination of moderators that might influence the effect size distribution.

In terms of possible bias, we found evidence that 14 effect sizes might have been missing from the *left* side of the distribution using both FE and RE models. Imputing these missing values decreased the estimated mean effect to d = 0.21 (95% CI = 0.13, 0.29) under FE and d = 0.25 (95% CI = 0.07, 0.42) under RE. Thus, for studies using a control-group design, estimates of the weighted average effect ranged from 0.21 to 0.25, when testing for possible bias. These estimates strengthen confidence that the effect of peer tutoring on academic achievement is positive and significantly different from zero. However, because the magnitude of the estimated effect size shrank by about half when accounting for possible bias, these estimates also suggest that the effect size of peer tutoring on academic achievement is likely to be small-to-medium in studies using a control-group design.

Table 3

Methodological Characteristic Moderator Analyses

	Pre-post Studies				Control-group Studies			
			95% CI				95% CI	
Moderator	Q_B / *k*	*d*	Low	High	Q_B / *k*	*d*	Low	High
Research design	n/a				42.06*** (19.47)***			
Pre-post	17	1.14*** (0.92)***	1.04 (0.63)	1.25 (1.22)	-			
Quasi-experimental	-				3	0.66*** (0.66)***	0.28 (0.28)	1.05 (1.05)
RCT/1 match	-				17	0.77*** (0.73)***	0.63 (0.56)	0.91 (0.89)
RCT/2 match	-				12	0.18** (0.23)***	0.06 (0.07)	0.29 (0.38)
Fidelity check	1.62 (0.07)				5.15* (2.22)			
Reported	14	1.16*** (0.90)***	1.05 (0.57)	1.27 (1.23)	22	0.38*** (0.44)***	0.29 (0.25)	0.48 (0.63)
Not reported	3	0.93*** (1.01)**	0.58 (0.31)	1.27 (1.71)	10	0.66*** (0.66)***	0.44 (0.44)	0.89 (0.89)
Outcome measure	90.82*** (90.82)***				0.80 (0.14)			
Standardized	9	0.52*** (0.52)***	0.34 (0.34)	0.70 (0.70)	18	0.39*** (0.47)***	0.29 (0.23)	0.50 (0.70)
Researcher	6	1.57*** (1.57)***	1.43 (1.43)	1.71 (1.71)	13	0.48*** (0.53)***	0.32 (0.33)	0.63 (0.72)
Unspecified	2	0.58** (0.58)**	0.19 (0.19)	0.97 (0.97)	1	0.77*** (0.77)***	0.04 (0.04)	1.49 (1.49)

Note. Fixed-effects estimates are presented outside parentheses and random-effects estimates are within parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

5.4 Moderator Analyses

Next, potential moderators of the effect of peer tutoring on academic achievement were examined. Grubb's outlier test was conducted on each moderator analysis effect size dataset, but detected no new outliers.

Methodology. Firstly, methodological characteristics thought to moderate peer tutoring effects were examined: these included research design, fidelity checks, and the outcome measures used. Table 3 presents the results.

Research design. For control-group studies, effect sizes varied with the type of research design (FE: $Q_B(2) = 42.06$, $p < .001$; RE: $Q_B(2) = 19.47$, $p < .01$), and a pairwise comparison revealed that control-group studies using a RCT/2 match design reported smaller effect sizes than studies using a RCT/1 match design (FE: $Q_B(1) = 40.47$, $p < .001$; RE: $Q_B(1) = 18.39$, $p < .001$). In order to double check that the low effect size for RCT/2 was not biased by subject we also compared math studies to literacy. For the RCT/2+ matched studies, there was no evidence that the $k = 8$ math effect sizes were significantly different from the $k = 4$ literacy effect sizes (FE: $Q_B(1) = 0.65$, $p = .42$; RE: $Q_B(1) < 0.01$, $p = .99$) and therefore, subject matter was not observed to have an impact on outcomes.

Also, the mean effect size of studies using a RCT 2+ match design was significantly smaller than the mean effect size of studies using a quasi-experimental. And the magnitude of the mean *gain* effect size associated with pre-post studies was also larger than the magnitude of mean *difference* effect size associated with all three control-group studies (see Table 3).

However, this finding should be interpreted with caution because mean gain and mean difference effect size statistics are not directly comparable (Lipsey & While 2011). Pairwise comparisons with control-group studies using a quasi-experimental design were not undertaken due to the small number of studies ($k = 3$) and the associated instability of the estimated average effect size.

Fidelity check. There was no evidence that pre-post studies conducting a fidelity check reported different effect sizes than studies not reporting a fidelity check. However, for control-group studies, effect sizes varied significantly depending on reporting a fidelity check under a FE model, $Q_B(1) = 5.15$, $p = .02$, but not under a RE model. According to the FE results, control-group studies conducting a fidelity check reported smaller effect sizes than

studies not reporting a fidelity check. However, the distinction should be interpreted with caution due to the small number of pre-post studies failing to report a fidelity check (k = 3).

Outcome measure. For pre-post studies, effect sizes varied depending on the type of outcome measure (FE: $Q_B(2) = 90.82$, $p < .001$; RE: $Q_B(2) = 90.82$, $p < .001$), and pairwise comparisons revealed that pre-post studies using researcher-modified measures reported larger effect sizes than those using standardized measures (FE: $Q_B(1) = 82.32$, $p < .001$; RE: $Q_B(1) = 82.32$, $p < .001$). Pairwise comparisons with the k = 2 pre-post studies using unspecified measures were not undertaken.

There was no evidence that effect sizes reported by control-group studies varied dependent on the type of outcome measures (FE: $Q_B(1) = 0.80$, $p = .37$; RE: $Q_B(1) = 0.14$, $p = .71$). This suggests that there was an association between the type of outcome measure and the magnitude of effect sizes for pre-post studies, but not for control-group studies.

Peer tutoring characteristics. Next, we examined peer tutoring characteristics thought to moderate effects: these characteristics included the type of peer tutoring, structure, and the use of a reward in the tutoring process. Table 4 presents the results.

Type of peer tutoring. For pre-post studies, effect sizes varied significantly with the type of peer tutoring (FE: $Q_B(2) = 54.21$, $p < .001$; RE: $Q_B(2) = 6.93$, $p < .05$). Pre-post studies using same-age reciprocal peer tutoring reported the largest effect sizes, followed by studies using cross-age non-reciprocal peer tutoring, but the difference between these effect sizes was only significant under a FE model, $Q_B(1) = 35.53$, $p < .001$, and not under a RE model, $Q_B(1) = 1.11$, $p = .29$. Pairwise comparisons were not undertaken with the one pre-post study using a mixed type of peer tutoring, or the k = 3 pre-post studies using same-age non-reciprocal peer tutoring.

For control-group studies, effect sizes also varied significantly with the type of peer tutoring (FE: $Q_B(2) = 28.85$, $p < .001$; RE: $Q_B(2) = 12.17$, $p < .01$). Similar to pre-post studies, control-group studies using same-age reciprocal peer tutoring reported the largest effect sizes, followed by studies using cross-age non-reciprocal peer tutoring. However, for control-group studies, the difference between these effect sizes was not significantly different (FE: $Q_B(1) = 0.75$, $p = .39$; RE: $Q_B(1) = 0.14$, $p = .71$). While this suggests that same-age reciprocal peer tutoring had a relatively greater effect on achievement compared to cross-age

non-reciprocal peer tutoring in pre-post studies compared to control-group studies, the distinction should be interpreted with caution due to the lack of robust model assumptions.

Table 4

Peer Tutoring Characteristic Moderator Analyses

	Pre-post Studies				Control-group Studies			
			95% CI				95% CI	
Moderator	Q_B / k	d	Low	High	Q_B / k	d	Low	High
Type of peer tutoring	54.21*** (6.93)*				28.85*** (12.17)**			
Same-age non-reciprocal	3	0.47** (0.47)*	0.15 (0.11)	0.80 (0.84)	4	0.04 (0.09)	-0.13 (-0.15)	0.21 (0.32)
Same-age reciprocal	6	1.45*** (1.17)***	1.31 (0.80)	1.58 (1.55)	15	0.60*** (0.56)***	0.47 (0.36)	0.72 (0.75)
Cross-age non-reciprocal	7	0.69*** (0.87)***	0.48 (0.47)	0.90 (1.28)	13	0.50*** (0.62)***	0.28 (0.36)	0.65 (0.87)
Mixed	1	1.07** (1.07)**	0.39 (0.39)	1.75 (1.75)	-			
Structure	n/a				n/a			
Structured	16	1.17** (0.92)***	1.06 (0.62)	1.27 (1.23)	31	0.42*** (0.52)***	0.34 (0.35)	0.51 (0.68)
Unstructured	-				1	0.50 (0.50)	-0.34 (-0.34)	1.34 (1.34)
Reward	72.75*** (19.32)***				0.01 (0.86)			
Reward	4	1.56*** (1.49)***	1.42 (1.24)	1.71 (1.74)	10	0.42*** (0.42)***	0.26 (0.26)	0.57 (0.53)
No reward	13	0.64*** (0.72)***	0.49 (0.49)	0.80 (0.96)	22	0.43*** (0.55)***	0.32 (0.32)	0.57 (0.77)

Note. Fixed-effects estimates are presented outside parentheses and random-effects estimates are within parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Similar to pre-post studies, control-group studies using same-age reciprocal peer tutoring also reported significantly larger effect sizes than studies using same-age non-reciprocal peer tutoring (FE: $Q_B(1) = 27.86$, $p < .001$; RE: $Q_B(1) = 9.35$, $p < .001$). Control-group studies using cross-age non-reciprocal peer tutoring reported larger effect sizes than studies using same-age non-reciprocal peer tutoring (FE: $Q_B(1) = 13.78$, $p < .001$; RE: $Q_B(1) = 9.18$, $p < .01$).

Structure. Moderator analyses for structure on the pre-post or control-group studies were not undertaken, because all but one used structured peer tutoring.

Reward. While pre-post studies using a reward incentive reported larger effect sizes than studies not using a reward (FE: $Q_B(1) = 72.75$, $p < .001$; RE: $Q_B(1) = 19.32$, $p < .001$), there was no such evidence for control-group studies (FE: $Q_B(1) < 0.01$, $p = .93$; RE: $Q_B(1) = 0.86$, $p = .35$). However, due to the small number of pre-post studies providing a reward ($k = 4$), this differential finding should be interpreted with caution.

Sample characteristics. Next, sample characteristics thought to moderate peer tutoring effects were examined: These were SES, minority percentage, grade level, and ability level representation in the sample. Table 5 presents the results.

SES. For pre-post studies, effect sizes varied significantly with students' SES (FE: $Q_B(2) = 66.21$, $p < .001$; RE: $Q_B(2) = 12.74$, $p < .001$). Pre-post studies involving students with low SES reported larger effect sizes than studies involving students with mixed SES (FE: $Q_B(1) = 22.25$, $p < .001$; RE: $Q_B(1) = 7.14$, $p < .01$), and studies failing to specify students' SES (FE: $Q_B(1) = 56.91$, $p < .001$; RE: $Q_B(1) = 9.95$, $p < .01$).

For control-group studies, effect sizes also varied significantly with students' SES under an FE model, though not under a RE model (FE: $Q_B(3) = 33.12$, $p < .001$). According to the FE models, control-group studies involving students with low SES reported the largest effect sizes, followed by studies involving students with average SES. Control-group studies involving students with low SES also reported larger effect sizes than studies involving students with mixed SES ($Q_B(1) = 31.07$, $p < .001$), and studies failing to specify students' SES ($Q_B(1) = 9.09$, $p < .01$).

Table 5

Sample Characteristic Moderator Analyses

	Pre-post Studies				Control-group Studies			
			95% CI				95% CI	
Moderator	Q_B / *k*	*d*	Low	High	Q_B / *k*	*d*	Low	High
SES	66.21*** (12.74)**				33.12*** (6.94)†			
Low	5	1.54*** (1.40)***	1.40 (1.12)	1.68 (1.68)	8	0.79*** (0.71)***	0.62 (0.46)	0.95 (0.97)
Average	1	0.22*** (0.22)***	-0.42 (-0.42)	0.85 (0.85)	3	0.70*** (0.70)***	0.31 (0.31)	1.10 (1.09)
High	1	1.68*** (1.68)***	0.76 (0.76)	2.61 (2.61)	-			
Mixed	3	0.68*** (0.71)**	0.35 (0.28)	1.01 (1.13)	14	0.20** (0.34)**	0.07 (0.13)	0.32 (0.55)
Unspecified	7	0.61*** (0.71)***	0.42 (0.39)	0.81 (1.04)	7	0.40** (0.39)**	0.19 (0.19)	0.59 (0.59)
Minority percentage	59.65*** (5.35)*				4.02* (0.80)			
≤ 50% sample	10	0.63*** (0.70)***	0.47 (0.45)	0.80 (0.95)	19	0.37*** (0.45)***	0.26 (0.23)	0.47 (0.67)
> 50% sample	7	1.48*** (1.22)***	1.34 (0.87)	1.61 (1.57)	11	0.58*** (0.58)***	0.39 (0.39)	0.77 (0.77)
Unspecified	-				2	0.72** (0.70)*	0.24 (0.13)	1.19 (1.26)
Grade level	8.56** (1.52)				0.01 (0.07)			

Elementary	13	1.19*** (0.98)***	1.08 (0.63)	1.31 (1.32)	30	0.42*** (0.51)***	0.34 (0.35)	0.51 (0.67)
High	4	0.68*** (0.68)***	0.35 (0.35)	1.00 (1.00)	2	0.42 (0.42)	-0.22 (-0.22)	1.07 (1.07)
Ability level	2.27 (0.03)				5.19* (2.81)†			
Low	5	1.29*** (0.85)**	1.10 (0.25)	1.47 (1.45)	10	0.54*** (0.54)***	0.36 (0.36)	0.71 (0.71)
Average	2	1.36*** (1.36)***	0.94 (0.94)	1.77 (1.77)	1	0.88** (0.88)**	0.22 (0.22)	1.53 (1.53)
Mixed	8	1.11*** (0.92)***	0.96 (0.42)	1.25 (1.42)	19	0.30*** (0.31)**	0.20 (0.10)	0.40 (0.51)
Unspecified	2	0.58** (0.58)**	0.19 (0.19)	0.97 (0.97)	-			

Note. Fixed-effects estimates are presented outside parentheses and random-effects estimates are within parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Minority percentage. Pre-post studies involving samples with ≤ 50% minority students reported smaller effect sizes than studies involving samples with > 50% minority students (FE: $Q_B(1) = 59.65$, $p < .001$; RE: $Q_B(1) = 5.35$, $p = .04$). Control-group study effect sizes varied significantly depending on the minority student percentage under a FE model, $Q_B(1) = 4.02$, $p < .05$. According to the FE results, control-group studies involving samples with ≤ 50% minority students reported smaller effect sizes than studies involving samples with > 50% minority students.

Grade level. For pre-post studies, effect sizes also varied significantly depending on students' educational level under a FE model, $Q_B(1) = 8.56$, $p < .01$. According to the FE results, pre-post studies involving elementary school students reported larger effect sizes than studies involving high school students.

Ability level. For pre-post studies, there was no evidence that studies involving students with low ability levels reported significantly different effect sizes than studies involving students with a mixed ability level.

There was evidence that control-group studies involving tutees with low ability levels reported larger effect sizes than studies involving tutees with a mixed ability level under a FE model, $Q_B(1) = 5.19$, $p = .02$, but not under a RE model.

Group characteristics. Next, two group characteristics thought to moderate peer tutoring effects were examined: gender composition and ability composition. Table 6 presents the results.

Gender composition. Moderator analyses for gender composition on the pre-post studies were not conducted, because all but one involved mixed-gender peer tutoring groups. However, control-group studies involving same-gender peers reported larger effect sizes than studies involving both mixed-gender peers (FE: $Q_B(1) = 9.63$, $p < .01$; RE: $Q_B(1) = 5.67$, $p = .01$). Pairwise comparisons with the one control-group study involving both same- and mixed gender peer tutoring groups were not conducted.

Ability composition. While the omnibus results indicated that pre-post effect sizes varied significantly depending on ability composition (FE: $Q_B(2) = 21.40$, $p < .001$; RE: $Q_B(2) = 12.58$, $p < .01$), pairwise comparisons were not conducted due to the small number of studies

(k = 2 studies involved same ability peers and k = 3 involved both mixed- and same-ability peers). For control-group studies, there was no evidence that studies involving mixed-ability peers reported different effect sizes than those involving same-ability peers (FE: $Q_B(1) < 0.01$, $p = .96$; RE: $Q_B(1) = 0.31$, $p = .57$). Pairwise comparisons with the one control-group study involving both mixed- and same-ability peers were not undertaken.

Intervention characteristics. Finally, intervention characteristics thought to moderate peer tutoring effects were explored: intervention length, training dosage, and setting. Table 7 presents the results.

Intervention length. For pre-post studies, effect sizes varied depending on the length of the intervention (FE: $Q_B(3) = 84.34$, $p < .001$; RE: $Q_B(3) = 55.02$, $p < .001$). Pre-post studies involving interventions lasting 13-52 weeks reported the largest effect sizes, followed closely by studies involving interventions lasting more than 52 weeks. But these estimates should be interpreted with caution due to the small number (i.e., $k < 3$) of pre-post studies contributing to the group effect size estimates. There was no evidence that pre-post studies involving interventions lasting 6 weeks reported different effect sizes than studies involving interventions lasting 7-12 weeks (FE: $Q_B(1) = 2.68$, $p = .10$; RE: $Q_B(1) = 1.78$, $p = .18$).

For control-group studies, effect sizes varied depending on the length of the intervention (FE: $Q_B(3) = 34.75$, $p < .001$; RE: $Q_B(3) = 11.53$, $p < .01$). Control-group studies with interventions lasting more than 52 weeks reported lower effect sizes than studies with interventions lasting 6 weeks (FE: $Q_B(1) = 6.07$, $p = .01$; RE: $Q_B(1) = 4.53$, $p = .03$), 7-12 weeks (FE: $Q_B(1) = 15.93$, $p < .001$; RE: $Q_B(1) = 9.68$, $p < .01$), and 13-52 weeks, (FE: $Q_B(1) = 28.67$, $p < .001$; RE: $Q_B(1) = 3.58$, $p = .05$). There was no evidence that control-group studies with interventions lasting 6 weeks reported different effect sizes than studies with interventions lasting 7-12 weeks or 13-52 weeks. Nor was there any evidence that control-group studies with interventions lasting 7-12 weeks reported different effect sizes than studies with interventions lasting 13-52 weeks. Thus, findings were similar for pre-post and control-group studies with the exception of the smaller effect sizes reported by control-group studies involving interventions lasting more than 52 weeks.

Training dosage. For pre-post studies, effect sizes also varied significantly depending on the training dosage (FE: $Q_B(2) = 82.76$, $p < .001$; RE: $Q_B(2) = 43.64$, $p < .001$). Pre-post studies involving ≥ 3 training sessions reported larger effect sizes than studies involving less than 3 training sessions (FE: $Q_B(1) = 73.10$, $p < .001$; RE: $Q_B(1) = 42.99$, $p < .001$), and

studies involving an unspecified number of training sessions (FE: $Q_B(1) = 29.62$, $p < .001$; RE: $Q_B(1) = 11.71$, $p < .01$). Pre-post studies involving an unspecified number of training sessions also reported larger effect sizes than studies involving less than 3 training sessions under a FE model, $Q_B(1) = 4.31$, $p = .04$, but not under a RE model, $Q_B(1) = 2.36$, $p = .12$.

For control-group studies, effect sizes varied significantly depending on training dosage under a FE model, $Q_B(2) = 24.42$, $p < .001$, but not under a RE model, $Q_B(2) = 2.63$, $p = .27$. Focusing on the FE results therefore, control-group studies involving an unspecified number of training sessions reported larger effect sizes than studies involving less than 3 training sessions, $Q_B(1) = 24.32$, $p < .001$, and studies involving ≥ 3 training sessions, $Q_B(1) = 9.44$, $p < .01$. There was no evidence that control-group studies involving ≥ 3 training reported different effect sizes than studies involving less than 3 training sessions. These findings suggest that training dosage was associated with differential effects in pre-post and control-group studies. Compared to studies with < 3 training sessions, studies with ≥ 3 training sessions were associated with increased effects in pre-post studies, but no significant difference in control-group studies.

Setting. Finally, pre-post effect sizes also varied significantly depending on setting under a FE model, $Q_B(2) = 49.05$, $p < .001$, but not under a RE model, $Q_B (2) = 2.42$, $p = .30$. Focusing on the FE results therefore, pairwise comparisons revealed that pre-post studies conducted in an intact classroom reported larger effect sizes than studies conducted in a general academic setting under a FE model, $Q_B(1) = 37.20$, $p < .001$. Pairwise comparisons were not undertaken with the $k = 3$ pre-post studies conducted in a laboratory setting.

Control-group effect sizes also varied significantly depending a study's setting (FE: $Q_B(2) = 30.91$, $p < .001$; RE: $Q_B (2) = 6.84$, $p < .05$), and pairwise comparisons revealed that control-group studies conducted in an intact classroom reported smaller effect sizes than studies conducted in a general academic setting (FE: $Q_B(1) = 29.74$, $p < .001$; RE: $Q_B(1) = 6.15$, $p = .01$). They also reported smaller effect sizes than studies conducted in a laboratory setting under a FE model, $Q_B(1) = 4.64$, $p = .03$, but not under a RE model. There was no evidence that control-group studies conducted in a general academic setting reported different effect sizes than studies conducted in a laboratory setting. While this suggests that setting was associated with differential effects in pre-post and control-group studies, the distinction should be interpreted with caution because the significant pre-post contrast was not robust to model assumptions.

Table 6.

Group Characteristic Moderator Analyses

	Pre-post Studies				Control-group Studies			
			95% CI				95% CI	
Moderator	Q_B / *k*	*d*	Low	High	Q_B / *k*	*d*	Low	High
Gender composition	0.83 (0.11)				9.63** (5.67)*			
Same	-				7	0.74*** (0.74)***	0.47 (0.47)	1.01 (1.01)
Mixed	16	1.14*** (0.93)***	1.04 (0.62)	1.25 (1.23)	24	0.28*** (0.37)***	0.19 (0.23)	0.38 (0.51)
Both same and mixed	1	0.79* (0.79)*	0.03 (0.03)	1.55 (1.55)	1	1.10*** (1.10)***	0.84 (0.84)	1.36 (1.36)
Ability composition	21.40*** (12.58)**				0.01 (0.31)			
Same	2	0.56** (0.57)***	0.19 (0.08)	0.94 (1.06)	3	0.40* (0.40)*	0.03 (0.05)	0.76 (0.76)
Mixed	12	1.06*** (0.87)***	0.92 (0.47)	0.94 (1.27)	28	0.42*** (0.52)***	0.32 (0.34)	0.51 (0.70)
Both mixed and same	3	1.46*** (1.43)***	1.27 (1.20)	1.65 (1.66)	1	0.56** (0.56)**	0.03 (0.20)	0.92 (0.92)

Note. Fixed-effects estimates are presented outside parentheses and random-effects estimates are within parentheses. $^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table 7
Intervention Characteristic Moderator Analyses

	Pre-post Studies		95% CI		Control-group Studies		95% CI	
Moderator	Q_B / *k*	*d*	Low	High	Q_B / *k*	*d*	Low	High
Intervention length	84.34*** (55.02)***				34.75*** (11.53)**			
6 weeks	5	0.76*** (0.82)***	0.51 (0.38)	1.01 (1.26)	6	0.63** (0.63)**	0.26 (0.26)	0.99 (0.99)
7-12 weeks	7	0.49*** (0.49)***	0.29 (0.29)	0.70 (0.70)	13	0.56*** (0.59)***	0.40 (0.42)	0.72 (0.84)
13-52 weeks	3	1.65*** (1.56)***	1.44 (1.18)	1.85 (1.94)	7	0.76*** (0.56)**	0.57 (0.21)	0.95 (0.91)
> 52 weeks	2	1.49*** (1.49)***	1.29 (1.29)	1.69 (1.69)	6	0.14* (0.17)†	0.01 (-0.02)	0.27 (0.37)
Training dosage	82.76*** (43.64)***				24.42*** (2.63)			
< 3 sessions	5	0.47*** (0.47)***	0.26 (0.30)	0.68 (0.68)	17	0.26*** (0.44)***	0.15 (0.23)	0.38 (0.65)
≥ 3 sessions	6	1.57*** (1.51)***	1.43 (1.28)	1.71 (1.74)	8	0.40** (0.40)***	0.21 (0.21)	0.58 (0.58)
Unspecified	6	0.80*** (0.78)***	0.57 (0.44)	1.04 (1.13)	7	0.78*** (0.63)***	0.60 (0.33)	0.96 (0.93)
Setting	49.05*** (2.42)				30.91*** (6.84)*			
General academic	7	0.68*** (0.81)***	0.48 (0.46)	0.88 (1.16)	14	0.77*** (0.67)***	0.61 (0.44)	0.92 (0.89)
Intact classroom	6	1.43*** (1.15)***	1.30 (0.76)	1.57 (1.54)	13	0.23*** (0.31)***	0.12 (0.14)	0.34 (0.48)
Laboratory	4	0.60*** (0.69)*	0.27 (0.16)	0.92 (1.23)	5	0.58*** (0.60)***	0.28 (0.26)	0.88 (0.93)

Note. Fixed-effects estimates are presented outside parentheses and random-effects estimates are within parentheses. †$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$.

6. Discussion

The aim of this meta-analysis was to investigate various topics such as: a) exploring the role of the experimental design in influencing the magnitude of the effect size for peer tutoring, b) whether author developed assessment instruments differed compared to instruments developed independently, c) the overall effect size of peer tutoring for attainment, and d) identifying the effect size for various established variables. The following were the findings:

The meta-analysis provided in this paper primarily maps *ES* of a homogenous intervention in a homogeneous population, peer tutoring in primary and secondary schools, as a function of experimental design. It was reported in a Best Evidence Encyclopaedia review that *ES* increased as experimental design moved from matched/clustered RCT to quasi-experimental design (Slavin, Lake, Hanley & Thurston, 2012). However, the conclusion of Slavin, et al, (2012) was not supported by a systematic approach that quantified the extent of this effect. The presented meta-analysis has now started to define the influence that experimental design can have on outcomes. While it was not possible to make any sensible comparison between the quasi-design studies and other designs due to the small number of quasi-design studies, the findings from this study suggest that the stricter the research design the lower the effect size is.

Regarding the influence of the research instrument the findings were mixed; with the researcher made/modified instruments showing a significantly larger mean effect size than standardised established instrument for the single group design studies, but not for the control group studies.

In terms of the overall effect size, the findings in this study supports previous meta-analysis in peer tutoring, presented in table 1, section 2.2. Although this study cannot directly be compared to that of Leung (2014) finding of *ES* *d*=0.26, which had a more heterogeneous population and looked at more diverse outcome measures, when looking at the controlled group studies after computing publication bias investigation we find similar results, with fixed model *ES* *d*=0.21 and random model *ES* *d*=0.25. Without publication bias investigations the *ES* ranged from *d*=42 (FE) and d=0.51 (RM). Thus it can be concluded that in peer tutoring the ES is small to moderate. This review and meta-analysis updates previous studies.

Regarding exploring established variables in peer tutoring literature it was possible to determine that peer tutoring still appears to provide positive benefits to students when delivered in elementary and high school settings. In terms of the nature of peer tutoring it was noted that same-age reciprocal peer tutoring appeared to provide greatest benefit to students followed by cross-age fixed role peer tutoring. Peer tutoring appeared to work best with students from low SES and in groups with a significant proportion of minority students. Although data was somewhat mixed there was some evidence that students from low attainment levels may benefit the most from implementing peer tutoring.

These findings have implications for future meta-analyses and trials in general:

Apart from publication and dissemination bias that effect the ES in meta-analysis (Rothstein, Sutton & Bornstein, 2005; Song, et al, 2000), this paper has illustrated that the ES and outcomes of meta-analysis could be influenced by the relative proportion of studies of a certain design, and using particular instrument types included in samples examined (Berge & Sandercock, 2002; Oakley, 2006). Hence, there is a possibility that when comparison of teaching methods takes place the conclusion is biased due to the design characteristics. Consequently, meta-analysts and those doing research-synthesis need to either a) develop an algorithm which deals with different designs (Rubin, 1993), b) ensure that all groups in each 'arm' of the meta-analysis are balanced in terms of design (Charlson & Schmidt, 1999), and, c) create criteria for inclusion standards that create a homogenous sample (Slavin, 2008a).

The aim should always be to include studies with only the strongest research designs, as emphasised by Slavin (2008a), considering that the purpose of meta-analysis is to find out what works from the best evidence.

The findings from this study justify the request to reduce bias in Social Science research. One of the ways previously suggested that may be successful at reducing bias would be to agree on some form of guidelines which can serve as a constitution for educational research that use RCT designs, hence clarifying the methods in conducting and reporting educational trials (Thurston, 2008). A model for such an approach exists in other research domains. For example CONSORT (CONSORT, 2010) in medical research. Two of the main CONSORT (2010) recommendations for conducting strong RCTs and protecting the community from researcher bias, are points 10 and 11a; i.e. that a) the random allocation of condition is made by a third independent party, b) the results are assessed blindly, and c) that there is no contact between those who allocated the conditions and those who blindly assessed the results.

Independent allocation to conditions in education is not a new idea, Puffer, Torgerson, & Watson, (2003) have already suggested that independent randomisation should be the practice in education. In an investigation of 76 meta-analyses in medicine and health 58% had blind assessment of results, and those trials without blind assessment had exaggerated the effects of the intervention by 7%. One of the main reasons for this characteristic according to CONSORT (2010) is that un-blinded researchers bias through particular strategies, i.e. selecting favourable time-points, outcomes, or even by removing participants from the analyses.

When it comes to blind assessment, CONSORT strongly emphasises that "Blinding is an important safeguard bias, particularly when assessing subjective outcomes". It is uncommon to see independent randomisation reported in educational trials, let alone blind assessment. Tymms, et al, (2008) are correct in criticising policymakers for downplaying the role of RCTs in education, wasting resources, not improving standards and gaining ill advice from the educational research community. On the other hand, it is reported that the falsification of RCTs is one of the main reasons why policymakers have ignored some of the most important research findings in education (Newman, 2008).

There is no doubt that researcher bias is one of the greatest threats to the validity of educational research, and it has been reported that some educational researchers strive for *'fame and glory'* (Newman, 2008). There is a positivist bias to published studies and a substantive amount of studies that did not achieve significant change remain unpublished (Rosenthal, 1979), in meta-analysis terms referred to as publication bias.

Finally, RCT designs should match control and experimental samples on more than one criterion at pre-test, *we recommend blind (at multi-levels) clustered RCTS with stratified random selection and allocation to conditions*. Hence, we call for a separation of research competences in education with independent teams assigning to condition, implementing an intervention, selecting outcome measures and undertaking analysis blind to conditions. The need for such separation of powers is paramount especially when considering that in Social Sciences evidence is not easily disproven; research can then be viewed with a degree of objectivity. Such transparency and division of competences would produce a high quality of researchers and research.

Acknowledgement

Special thanks to Antje Hornburg, Head of Education, NeuroPartners.

References

Berge, E., & Sandercock, P. (2002). 'The nuts and bolts of doing a clinical trial', in Duley, L., Farrell, L., & Farrell, B. (eds.*) Clinical Trials* (Chapter 7). London: BMJ Publishing. 72-80.

Bowman-Perrot, L., Davis, H., Vannest, K., Williams, L. (2013). Academic Benefits of Peer Tutoring: A meat-Analytic of Singe-Case Research. School Psychology Review, Volume, (1), pp.39-55.

Campbell, D. T & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Boston: Houghton Mifflin Company.

Carlson, K. D., & Schmidt, F. (1999). Impact of experimental design on effect size: Findings from the research literature on training, *Journal of Applied Psychology*, 84 (6), 851-862.

Coe, R. (2002). It's the effect size stupid: What effect size is and why it is important. *Paper presented at the British Educational Research Association annual conference*, Exeter, 12-14, September, 2002.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal,19*, 237-248.

Cooper, H.M & Hedges, L.V (eds.) (1994). *The handbook of research synthesis*. New York: The Russell Sage Foundation.

Cook, S., Scruggs, T., Mastropieri, M., & Casto, G. (1986). Handicapped Students as tutors. Journal of Special Education, 19(4), 483-492.

Fantuzzo, J. W., King, J. A., & Heller, L. Rio. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology*, 84(3), 331-339.

Fern, E.F, & Monroe, K. B. (1996), Effect-Size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, 23(2), 89-105.

Fitz-Gibbon, C.T. (1990). Success and failure in peer tutoring experiments. In

Goodlad, S., & Hirst, B. (Eds.). *Exploration in Peer Tutoring*. Oxford: Basil Blackwell.

Ginsburg-Block. M. D., Rohrbeck, C.A., & Fantuzzo J.W., (2006). A Meta-Analytic Review of Social, Self-Concept, and Behavioural Outcomes of Peer-Assisted Learning. *Journal Of Educational Psychology*, 98 (4), pp.732-749.

Ginsburg-Block. M., & Fantuzzo, J. (1997). Reciprocal Peer Tutoring: An analysis of "teacher" and "student" interactions as a function of training and experience. *School Psychology Quarterly*, 12(2), 134-149.

Ginsburg-Block, M. D., & Fantuzzo, J.W. (1998). An evaluation of the relative effectiveness of NCTM standards-based Interventions for low-achieving urban elementary students. *Journal of Educational Psychology*, 90 (3), 560-569.

Greenwood, C. R, Delquadri, J. C., & Hall, R. V. (1989). Longitudinal effects of class wide peer tutoring. *Journal of Educational Psychology*, 81, (3), 371-383.

Greenwood, C.R., Dinwiddie, G., Bailey, V., Carta, J.J., Dorsey, D., Kohler, F.W., Nelson, C., Rotholz, D., & Schulte, D. (1987). Field replication of class wide peer tutoring. *Journal of Applied Behaviour Analysis*, 20 (2), 151-160.

Hammersley, M. (1997). Educational research and teaching: A response to David Hargreaves' TTA lecture. *British Educational Research Journal, 23* (2). 141-161.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107-128.

Higgins, S., Kokotsaki, D., & Coe, R.J. (2011). Toolkit of strategies to improve learning: Summary for schools spending the pupil premium. Sutton Trust: London. Retrieved on the 12th June 2012, from: http://www.suttontrust.com/research/toolkit-of-strategies-to-improve-learning/

Hodkinson, P. (2000). The contested field of educational research: The new research orthodoxy and the limits of objectivity. *Paper presented at the British Educational Research Association Annual Conference*, University of Leeds. Retrieved November, 16th , 2011, from: http://www.leeds.ac.uk.educol/documents/00001875.doc.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81* (396), 945-960.

Lachin, J.M., Matts, J.P., & Wei, L.J. (1988). Randomization in clinical trials: Conclusions and recommendations. *Control Clinical Trials, 9* (4), 365–74.

Lemons, C.J., Fuchs, D., Gilbert, J., Fuchs, L.S (2014). Evidence-Based Practice in a

Changing World: Reconsidering the Counterfactual in Education Research. *Educational Researcher*. 43(5), pp 242-252.

Leung, K, C. (2014). Preliminary Empirical Model of Crucial Determinant of Best Practice for Peer Tutoring on Academic Achievement. Journal Of Educational Psychology. http://dx.doi.org/10.1037/a0037698

Lipsey, M. S., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioural treatment. *American Psychologist*, 48 (12), 1181-1209.

Lipsey, M. W., & Wilson, B. D. (2001). *Practical Meta-Analysis.* Sage Publications Ltd: London.

Mathes, P, and Fuch, L (1994). The efficacy of peer tutoring in reading strategies for students with mild disabilities: A best-evidence synthesis. School Psychology Review. 23(1), 59-80.

McKinstery, J., & Topping, K. (2003). Cross-age peer tutoring of thinking. *Educational Psychology in Practice: theory, research and practice in educational psychology*, 19(3),199-217.

Merrett, F., & Mottram, S.(1997). Do Boys or Girls Make Better Reading Tutors? An Empirical Study to Examine Children's Effectiveness as Tutors Using the Pause, Prompt and Praise Procedures, *Educational Psychology: An International Journal of Experimental Educational Psychology*, Vol. 17 (4), pp. 419-432, DOI: 10.1080/0144341970170404

Miller, D., Topping, K., & Thurston, A. (2010). Peer tutoring in reading: The effects of role and organization on two dimensions of self-esteem. *British Journal of Educational Psychology*, 80(3), 417–433.

Morrison, K. (2001). Randomised controlled trials for evidence-based education: some problems in judging what works. *Evaluation and Research in Education, 15* (2), 69-83.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M., & Altman, D.G., for the CONSORT Group. (2010). Explanation and elaboration: updated guidelines for reporting parallel group randomised trial. *BMJ*.340:c869.

Newman, M. (2008). High quality randomized experimental research evidence: necessary but not sufficient for effective education policy. *The Psychology of Education Review, 32* (2), 14-16.

Oancea, A. (2005). Criticisms of educational research: key topics and levels of analysis.

British Educational Research Journal, *31* (2), 157-183.

Oakley, A. (2006). Resistances to 'new' technologies of evaluation: education research in the UK as a case study. *Evidence and Policy*, *2* (1), 63-87.

Puffer, S., Torgerson, D.J. & Watson, J. (2003). Evidence for risk of bias in cluster randomized trials: Review of recent trials published in three general medical journals. *British Medical Journal*, 327(7418), 785–789.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 95*, 240-257.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin. 86(3), 638–641.

Roseth, C., Johnson, D.W. & Johnson, R.T. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin*, *134* (2), 223-246.

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments.* Chichester, UK: Wiley.

Rubin, D.B. (1993). Statistical Tools for Meta-Analysis: From Straightforward to Esoteric. In *Interpersonal Expectations: Theory, Research, and Application*. P.D. Blanck (ed.). Cambridge University Press, pp. 400-417.

Ruiz-Primo, M.A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Slavin, R. E. (2008a). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14.

Slavin, R.E. (2008b). Evidence-based Reform. Which Evidence Counts? *Educational Researcher*, 37 (1), 47-50.

Slavin, R. E., & Madden. N. A. (2008). Understanding bias due to measures inherent to treatments in systematic reviews in Education. Retrieved July10th, 2012, From: http://www.bestevidence.org/methods/understand_bias_Mar_2008.pdf.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective program for middle and high school reading: A best-evidence synthesis. *Reading Research Quarterly*, 43 (3). 290-322.

Slavin, R.E., Groff, C., & Lake, C. (2009). Effective program for middle and high school

mathematics: A Best-Evidence synthesis. *Review of Educational Research*, 79 (2), 839-911.

Slavin, R., Lake, C., Hanley, P., & Thurston, A. (2012). *Effective programs for elementary science: A Best-Evidence synthesis.* Johns Hopkins University: Baltimore, USA.

Slavin, R.E., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370-380.

Song, F., Easterwood, A., Gilbody, S., Duley, L., & Sutton, A.J. (2000). Publication and other selection biases in systematic reviews. *Health Technology Assessment*, 4(10). 1-115.

Thurston, A. (2008). Cluster randomised controlled trials: The way forward for educational research? *The Psychology of Education Review, 32* (2), 21-23.

Thurston, A., Burns, V., Topping, K.J., & Thurston, M.J. (2012). Social effects of peer tutoring. *American Educational Research Association Annual Gathering Vancouver,* 10th-14th April 2012.

Thurston, A., Duran, D., Cuningham, E., Blanch, S., & Topping, K. (2009). International online reciprocal peer tutoring to promote modern language development in primary schools. *Computers & Education,* 53 (2), 462–472.

Thurston, A., MacNia, S., & Keenan, C. (2015). Differential impact of peer tutoring in mathematics in English and Irish Medium elementary schools. *American Educational Research Association Annual Gathering, Chicago,* 16th-20th April 2015.

Thurston, A., & Topping, K.J. (2008) A randomized trial of paired reading in elementary schools. *American Education Research Association Annual Gathering*. New York, USA, 24th-30th March 2008.

Topping, K.J., & Thurston, A. (2008). A randomized trial of maths tutoring in elementary schools. *American Education Research Association Annual Gathering*. New York, USA, 24th-30th March 2008.

Topping, K., & Ehly, S. (1998). Introduction to peer learning. In Topping, K.J., & Ehly (eds.). *Peer Assisted Learning* (pp.1-25). London UK: Lawrence Erlbaum.

Topping, K., Campbell. D., Walter, S., & Andrea, J. (2003). Cross-age peer tutoring in mathematics with seven - and 11-year-olds: influence on mathematical vocabulary, strategic dialogue and self-concept. *Educational Research*, 45(3), 287-308.

Topping, K., Peter, C., Stephen, P., & Whale, M. (2004). Cross-age peer tutoring of science

in the primary school: influence on scientific language and thinking. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 24 (1), 57-75.

Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, *49* (3), 316-329.

Traianou, A., & Hammersley, M. (2008). Making science education evidence-based? Reflections on a teaching and learning research programme (TLRP) study. *Oxford Review of Education*, *34* (4), 461–481.

Trochim, W. (2012). Design, the research method knowledge base. Retrieved January 14th , 2012, from: http://www.socialresearchmethods.net/kb/design.php.

Tymms, P.B., Merrell, C., & Coe, R.J. (2008). Educational policies and randomized controlled trials. *The Psychology of Education Review, 32* (2), 3-7 & 26-29.

Tymms, P., Merrell, C., Thurston, A., Andor, J., Topping, K.J., & Miller, D.J.(2011). Improving attainment across a whole district: Peer tutoring in a randomised controlled trial. *School Effectiveness and School Improvement, 22* (3), 265-289.

Wright, J. and Cleary, K. S. (2006). Kids in the tutor seat: Building schools' capacity to help struggling readers through a cross-age peer-tutoring program. *Psychol. Schs*., 43: 99–107.

Appendix

(A) Study Characteristics

We coded seven broad categories of study characteristics: (i) the research report, (ii) methodological, (iii) the type of peer tutoring, (iv) the sample, (v) the composition of the peer tutoring group, (vi) the intervention, (vii) the estimate of the effect of peer tutoring on academic achievement. Table 8 provides a complete list of coded study characteristics.

Table 8

Complete List of Coded Study Characteristics

i) Research report

1. Author name
2. Sample size
3. Decade published

ii) Methodological

1. Research design (pre-post, quasi-experimental, RCT 1 match, RCT 2+ match)
2. Fidelity check (reported, not reported)
3. Outcome measure (standardized, researcher modified)

iii) Peer tutoring

1. Subject matter
2. Peer tutoring type (same-age non-reciprocal, same-age reciprocal, cross-age, mixed)
3. Structure (structured, unstructured)
4. Reward incentive (tangible, nontangible)

iv) Sample

1. Nationality
2. Socioeconomic status (SES) (low, average, mixed, unspecified)
3. Minority percentage (≤ 50% sample, > 50% sample)
4. Grade level (elementary, high)
5. Ability level (low, average, mixed, unspecified)

v) Group composition

1. Gender (same, mixed, both same and mixed)
2. Ability (same, mixed, both same and mixed)

vi) Intervention

1. Length (weeks)
2. Training dosage (< 3 sessions, ≥ 3 sessions)
3. Setting (laboratory, school)

vii) Estimate of the effect

1. Direction of the effect
2. Magnitude of the effect

Coder reliability. Two graduate students extracted information from articles. The first author extracted information from all articles, and the second graduate student extracted information from 15 randomly selected articles from the final sample of 41, which corresponds to 37% of the population. Inter-coder agreement exceeded 90% for all elements. The first and third authors independently computed all effect size estimates. All disagreements were resolved through discussion with the second or third author. This process of ensuring high reliability is well-established in the literature (e.g., Cooper & Hedges, 1994; Lipsey & Wilson, 2001; Rosenthal, 1987; Roseth et al., 2008);

Effect size estimation. We used the standardized mean difference to estimate the effect of peer tutoring (i.e., the d-index; Cohen, 1988), subtracting the mean of the control condition (or pre-test condition) from that of the peer tutoring condition and dividing the difference by the average of their standard deviations. If available, we calculated subgroup effect sizes based on the means, standard deviations, and sample sizes for outcome indicators. When sample size information was not available, we estimated the effect size from the means and standard deviations and corresponding inference test. When no sample sizes or inference tests were available, we assumed group sample sizes to be equal and estimated the effect size from the means and standard deviations. Finally, when means and standard deviations were not available, we estimated the effect size from the reported inferential statistics. All estimates of effect size, variance, and 95% confidence intervals (CIs) were calculated using the Comprehensive Meta-Analysis (CMA) statistical software package (Version 3.3; Borenstein, Hedges, Higgins, & Rothstein, 2014). We used Hedges *g* to correct for positive bias with small samples (Hedges & Olkin, 1985).

(B) Data Integration

Before integrating the effect sizes, we first counted the number of positive and negative effects. Then, for each research design category, we examined the distribution of effect sizes for statistical outliers. If outliers were identified using the test of Grubbs (1950), then effect size values were set at the value of the next nearest value.

We used Duval and Tweedie's (2000a, 2000b) trim-and-fill procedure to test whether the distribution of effect sizes might be biased by our sampling procedures or by our inclusion and exclusion criteria. This trim-and-fill procedure involved removing extreme effect size values based on small studies from the positive side of the funnel plot and imputing estimated values to approximate a normal distribution.

Calculating average effect sizes. Following recommended practice (e.g., Cooper & Hedges, 1994; Lipsey & Wilson, 2001), we calculated the average weighted effect sizes while correcting for the upward bias associated with small sample sizes. Specifically, we weighted each effect size by multiplying each independent effect by the inverse of its variance, then dividing the sum of these products by the sum of the inverses. We corrected for small sample bias using Hedges' g (Hedges & Olkin, 1985).

Shifting unit of analysis. To identify independent effect sizes, we used a shifting unit of analysis approach (Cooper, 1998). This means that separate effect sizes for each subgroup within a single study (e.g., high, medium, and low ability) were used when evaluating the moderating effect of subgroup membership (e.g., the moderating effect of ability level on peer tutoring's effect on achievement). When estimating peer tutoring's overall effect, however, we used the average of subgroup effect sizes so that only one effect size was used for each study.

Testing moderator effects. We tested possible moderators of peer tutoring effects using homogeneity analyses (Cooper & Hedges, 1994; Hedges & Olkin, 1985), whereby a significant Q_W statistic indicates a heterogeneous distribution of effect sizes, the I^2 index describes the extent of heterogeneity (Higgins & Thompson, 2002; Juedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006), and a significant Q_B statistic indicates that effect sizes differ between subgroups.

We conducted all homogeneity analyses twice, once using a fixed-effects model and once using a random-effects model. This approach allowed us to examine the sensitivity of our

analyses to different assumptions about the sources of error (e.g., sampling error, study-level error; see Greenhouse & Iyengar, 1994; Hedges & Vevea, 1998). We conducted all integration analyses using the CMA statistical software package (Version 3.3; Borenstein et al., 2014).